DOCUMENT RESUME

ED 397 105 TM 025 233

AUTHOR Samejima, Fumiko
TITLE Polychotomous Responses and the Test Score.
PUB DATE Apr 96
NOTE 20p.; Paper presented at the Annual Meeting of the National Council of Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; *Attitude Measures; Estimation (Mathematics); Models; *Responses; *Scores; Scoring; Test Interpretation; *Test Items
IDENTIFIERS Graded Response Model; Number Right Scoring; *Polychotomous Scoring

ABSTRACT

Traditionally, the test score represented by the number of items answered correctly was taken as an indicator of the examinee's ability level. Researchers still tend to think that the number-correct score is a way of ordering individuals with respect to the latent trait. The objective of this study is to depict the benefits of using ability estimates obtained directly from individuals' response patterns instead of their test scores, especially when responses are graded polychotomously. The importance of substantive model validation is also discussed. Mathematical models are presented to show that the use of the test score instead of the response pattern itself in ability or attitude estimation will, in general, reduce the accuracy of estimation. The loss of accuracy can be especially important when items are scored polychotomously. It is suggested that ability or attitude estimation be made from the response pattern using basic functions developed by F. Samejima in conjunction with the graded response model. In doing so, substantive model validation is essential. (Contains 2 figures, 1 table, and 13 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
FUMIKO SAMEJIMA
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 397 105

POLYCHOTOMOUS RESPONSES AND THE TEST SCORE[1]

Fumiko Samejima

The University of Tennessee

[1] Requests for reprints should be sent to Fumiko Samejima, Department of Psychology, 405 Austin Peay Bldg., University of Tennessee, Knoxville, Tennessee 37996-0900. E-mail: samejima@psych1.psych.utk.edu

TM025233

BEST COPY AVAILABLE

Traditionally, even before mental test theory was proposed, the test score represented by the number of items that are answered correctly was taken as an indicator of the examinee's ability level. Today, researchers still tend to think that the number-correct test score is a way of ordering individuals with respect to the latent trait, justifying it by saying that the test score is a *consistent* estimator of ability. This idea has gone far enough to use *monotone likelihood ratio* (e.g., Grayson, 1988; Hyunh, 1994) of the test score in the latent trait as a way of evaluating mathematical models.

Samejima (1969, 1972) proposed the general framework of ordered polychotomous responses, and Bock (1972) proposed a nominal model for non-ordered polychotomous responses which uncovers implicit orders among the nominal responses. After a couple of decades, aided by advanced computer technologies, researchers started using mathematical models that belong to these two big families of models for their research and also for more practical purposes of ability and attitude assessments. The idea of monotone likelihood ratio has also been extended to evaluate polychotomous response models.

In reality, there is no such things as an infinitely long test, and our objective should be to make the best use of information provided by a finite number of items in the test and estimate the individual's ability level as accurate as possible. Samejima (1969) has shown that, in general, the use of the test score, or any aggregate of response patterns, will decrease the amount of test information, which results in loss of accuracy in ability estimation.

An advantage of the use of polychotomous responses over the use of dichotomous responses lies in increment in the amount of test information. Generally speaking, however, loss of test information by using the test score, and hence loss of accuracy in ability estimation, is greater when responses are graded polychotomously.

The objective of the present study is to depict benefits of using ability estimates

obtained directly from individuals' response patterns instead of their test scores, especially when responses are graded polychotomously, and also discuss the importance of *substantive* model validation.

I. Rationale

Let θ be the latent variable which assumes any real number, g $(= 1, 2, ..., n)$ denote an item, K_g be a polychotomous response to item g and and k_g denote its realization. The operating characteristic, $P_{k_g}(\theta)$, of the discrete item response k_g is defined as

$$P_{k_g}(\theta) \equiv prob.\,[K_g = k_g \mid \theta] \ . \tag{1}$$

When ordered polychotomous responses are dealt with, K_g and k_g in (1) are replaced by the graded responses X_g and x_g $(= 0, 1, 2, ..., m_g)$, respectively.

Samejima (1973) defined the item response information function, $I_{k_g}(\theta)$, such that

$$I_{k_g}(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_{k_g}(\theta) \ , \tag{2}$$

and the item information function $I_g(\theta)$ was defined as its conditional expectation, given θ , of the item response information function

$$I_g(\theta) = E[I_{k_g}(\theta) \mid \theta] = \sum_{k_g} I_{k_g}(\theta)\, P_{k_g}(\theta) \ . \tag{3}$$

Equation (3) includes Birnbaum's item information function for a dichotomous item (Birnbaum, 1968) as a special case.

Let V be a response pattern, or a sequence of item responses, and v denote its realization. Assuming *local independence* (Lord & Novick, 1968), the operating characteristic, $P_v(\theta)$, of a specific response pattern v is given by

$$P_v(\theta) \equiv prob.\,[V = v \mid \theta] = \prod_{k_g \in v} P_{k_g}(\theta) \ . \tag{4}$$

Thus from (2) and (4) the response pattern information function, $I_v(\theta)$, can be written as

$$I_v(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_v(\theta) = \sum_{k_g \in v} I_{k_g}(\theta) \quad , \tag{5}$$

and the test information function, $I(\theta)$, is defined as the conditional expectation of the response pattern information function, given θ, and from (2), (3), (4) and (5)

$$I(\theta) = E[I_v(\theta) \mid \theta] = \sum_v I_v(\theta)\, P_v(\theta) = \sum_{g=1}^{n} I_g(\theta) \quad . \tag{6}$$

This amount of test information, or its square root, is used as a *local* measure of accuracy in ability or attitude estimation, when the response pattern is used as the basis of estimation.

Let T be any *aggregation* of response patterns, and t be its realization. The operating characteristic, $P_t(\theta)$, of a specific aggregation t is given by

$$P_t(\theta) = \sum_{v \in t} P_v(\theta) \quad . \tag{7}$$

When the aggregates t's are disjoint and each response pattern v belongs to one and only one t, as is the case with the test score, the test information function, $I^*(\theta)$, based on such an aggregation is defined by

$$I^*(\theta) = E[-\frac{\partial^2}{\partial\theta^2} \log P_t(\theta) \mid \theta] = E[\{\frac{\partial}{\partial\theta} \log P_t(\theta) \mid \theta\}^2] \quad , \tag{8}$$

as opposed to the original test information function, which can be written as

$$I(\theta) = E[-\frac{\partial^2}{\partial\theta^2} \log P_v(\theta) \mid \theta] = E[\{\frac{\partial}{\partial\theta} \log P_v(\theta) \mid \theta\}^2] \quad ,$$

which is obvious from (5) and (6). Using Cauchy-Schwarz's inequality, Samejima (1969) showed that

$$I^*(\theta) \leq I(\theta) \quad . \tag{9}$$

Inequality (9) implies that, if polychotomous response categories are more finely classified, as exemplified by a 7-point scale versus a 3-point scale, then in general the amount

of test information will be increased. More importantly, since the test score is an aggregation of response patterns, except for certain special cases, the amount of test information will be decreased if the test score is used as the basis of ability or attitude estimation, instead of the response pattern itself: the fact that leads to *inaccuracy* in the latent trait estimation.

Samejima (1969, 1972) proposed a sufficient condition for an item response to provide a unique local or terminal maximum likelihood estimate for any response pattern consisting of such item responses. Let $A_{k_g}(\theta)$ be such that

$$A_{k_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{k_g}(\theta) \quad . \tag{10}$$

The sufficient condition is that $A_{k_g}(\theta)$ be strictly decreasing in θ with non-negative and non-positive values for its two asymptotes, respectively. For brevity, this condition has often been called the *unique maximum condition*. It is noted from (2) and (3) that the first part of this condition can be rephrased, that is, the item response information function be positive for all θ except, at most, at an enumerable number of points where it is zero.

It has been shown (Samejima, 1969, 1972) that both the normal ogive model and the logistic model on the dichotomous response level, and also those on the polychotomous response level, satisfy this condition, and so does Bock's nominal response model. Thus in these models the likelihood function based on the response pattern, which is the same as $P_v(\theta)$ in (4), has a unique local or terminal maximum for every $v \in V$. Note, however, that neither the three-parameter logistic model nor the the three-parameter normal ogive model satisfies the unique maximum condition (Samejima, 1972, 1973) and thus for some response patterns the likelihood functions have multi-modes. Yen, Burket & Sykes (1991) have shown that multi-modality of the likelihood function occurs not infrequently for response patterns that usually come across in empirical data if the

three-parameter logistic model is used.

If an aggregate T exemplified by the test score is used instead of the response pattern V as the basis of ability or attitude estimation, however, unimodality of the likelihood function is *not* assured, even if the model itself satisfies the unique maximum condition. This is obvious from the likelihood function $P_t(\theta)$ shown in (7).

From (4) and (10) it is obvious that the likelihood equation is given by

$$\frac{\partial}{\partial\theta}\log P_v(\theta) \;=\; \sum_{k_g \in v} A_{k_g}(\theta) \;\equiv\; 0 \;. \tag{11}$$

Thus if the model satisfies the unique maximum condition the argorithm for obtaining the maximum likelihood estimat $\hat{\theta}_v$ will be easy and straightforward, using the basic functions. In this way, without usning any aggregates such as the test score, the examinee's ability can be estimated directly from his/her response pattern.

II. Test Score with Polychotomous Responses

In Rasch model (Rasch, 1960), equality in (9) holds if the number-right test score is used as T, since it is a sufficient statistic. Thus in this model accuracy in ability or attitude estimation will not decrease even if it is based on the number-right test score. It has been reported by many researchers, however, that in many cases estimated operating characteristics of the correct answer, or item characteristic curves, provide distinct discrimination powers for separate items in a test. Thus it will be an unjustified approach to adopt Rasch model in such cases even though the model porvides the benefit of mathematical simplicity.

When the discrimination parameter is considered as in the normal ogive and logistic models, the number-right test score is no longer a sufficient statistic. Thus inequality holds in (9) that results in the loss of accuracy in estimating the latent trait when the test score is used as the basis of ability or attitude estimation instead of the response

pattern itself. Decrease in the amount of test information is illustrated in Figure 1 for 10 hypothetical, dichotomous test items following the logistic model, whose operating characteristic, $P_g(\theta)$, of the correct answer to item g is given by

$$P_g(\theta) = \frac{1}{1+\exp[Da_g(\theta - b_g)]} \quad , \tag{12}$$

where a_g and b_g are the discrimination and difficulty parameters, respectively, and D is a scaling factor approximately equals 1.7 (Birnbaum, 1968). The values of these parameters for the 10 items are shown in the same figure. Considering that the number-right test score classifies individuals into only $(n+1)$ categories, whereas the number of possible response patterns is as large as 2^n, indicating that if $n = 10$ these numbers are 11 versus 1,024, decrease in the amount of test information is readily acceptable.

Insert Figure 1 About Here

When each item is polychotomously scored, the test score is given by $\sum_{x_g \in v} x_g$, where x_g $(= 0, 1, ..., m_g)$ is the graded item score. Thus even if the smallest number, $m_g = 2$, applies for each and every item, the number of possible response patterns becomes as large as 3^n, whereas that of distinct test scores is only $2n+1$. This implies that if $n = 10$ there are 59,049 possible response patterns in contrast to the 21 test score categories, and, for example, the test score 10 includes as many as 8,953 different response patterns. To be more specific, consider two response patterns, $(2,2,2,2,2,0,0,0,0,0)$ and $(0,0,0,0,1,1,2,2,2,2)$, with the 10 items arranged in the ascending order of difficulty. Should these two response patterns be treated as *equivalent* because they provide the same test score 10 ? Common sense tells us that these two response patterns that belong to the same test score category may represent

substantially different latent trait levels. Use of the test score results in the loss of all specific information contained in each response pattern, however, and these two response patterns are treated as if they were equal.

Insert Figure 2 About Here

Figure 2 illustrates decrease in the amount of test information caused by the use of the test score instead of the response pattern itself, for 10 hypothetical, ordered polychotomous items with $m_g = 2$ each, with the second difficulty parameter 0.5 higher than the first one for each dichotomous item illustrated in Figure 1. As is expected, a greater amount of decrement is seen in Figure 2 than in Figure 1.

A strength of the general ordered polychotomous model is that m_g can be different for separate items. In such a case, a single test score contains a greater variety of response patterns, and a greater loss of test information is expected. Another strength of the general ordered polychotomous response model lies in a variety of configurations of item response difficulty parameters within each item and across separate items. For example, the distance between the difficulty parameters between $x_g = 1$ and $x_g = 2$ may be substantially larger than the one between $x_g = 2$ and $x_g = 3$ within a single item, and this relationship may be reversed across two separate items. The test score *ignores* such distinct configurations of difficulty parameters, and treat them as if they were equal, whereas the likelihood equation (11) based on the response pattern reflects such differences as they are.

Scatter diagrams of the test score for polychotomous response items and an estimate of the individual's latent trait based on the response pattern usually provide a wide range of dispersion, especially when m_g is large for most items. This indicates that

the test score for polychotomous response items does not represent individual differences with respect to the latent trait.

It is easily seen from (11) that, as long as the model satisfies the unique maximum condition, for a test of n graded items, there are only $\sum_{g=1}^{n} m_g + n$ basic functions. From this small number of basic functions, using an appropriate, simple algorithm, $\prod_{g=1}^{n}(m_g + 1)$ likelihood equations can be produced, and the maximum likelihood estimate, $\hat{\theta}_v$, of the latent trait is obtained as the solution of the likelihood equation (11) for a specific response pattern v . For example, when $n = 10$ and $m_g = 2$ for all items, there are only 30 basic functions in total, and they provide $\hat{\theta}_v$ for each of 59,049 response patterns.

III. Substantive Criteria for Model Validation

Most researchers are concerned with the goodness of fit of the operating characteristics that a mathematical model provides to their set of empirical data as the criterion for model validation. Samejima (in press) pointed out the danger of this conventional way of model validation, illustrating that two mathematical models that are based on totally different rationale can produce practically identical sets of operating characteristics, and there will not be any way to decide which model is more suitable if curve fitting is exclusively used for model validation. In the same token monotone likelihood ratio of the test score cannot be used as a criterion for model validation, since the test score only asymptotically converges in probability to the latent trait.

Samejima (in press) proposed more substantive criteria for model validation such as the fitness of the rationale or principle behind the model to our data, additivity intrinsic in the model, its natural generalizability to a continuous model, satisfaction of the unique maximum condition, and orderliness of the modal points of the operating characteristics that the model provides. Among others, the rationale on which a specified mathematical

BEST COPY AVAILABLE

model is based should be considered intensively.

One of the issues that needs to be considered is how to accommodate differences in difficulty levels of items in ability estimation. On the dichotomous response level, for example, should the correct answer to a more difficult item be credited more than that to an easier item, or should failure in answering correctly to an easier item be penalized more than failure in answering a more difficult item correctly? Samejima (1969) pointed out that symmetric models that provide point-symmetric item characteristic curves with $(b_g, 0.5)$ as the point of symmetry, which are exemplified by the normal ogive and logistic models, have a certain logical problem, because they treat correct answers and incorrect answers symmetrically. accommodating both of the above philosophies. Taking the normal ogive model as an example, whose operating characteristic of the correct answer is given by

$$P_g(\theta) = \int_{-\infty}^{a_g(\theta - b_g)} \exp[-\frac{u^2}{2}] \, du \tag{13}$$

where a_g and b_g are the discrimination and difficulty parameters, respectively, this problem was presented and discussed in the original literature (Samejima, 1969, pages 85-86) as follows:

"Now let us consider a situation in which examinees are required to solve n dichotomous items, all of whose discriminating powers are the same, but whose difficulty levels are different from one another. For simplicity, let us suppose that $n = 5$ and the five items are denoted by 1, 2, 3, 4, and 5, in the order of easiness. If there are five examinees who have tried to solve all the five items but succeeded in only one, and each item solved is different from each other, to which of the five examinees should be assigned the highest value of estimate? The answer to this question may largely depend upon subjective judgments or preferences. On the normal ogive model, however, the answer is definite, as we can easily observe by following the preceding reasoning. We

can arrange the five response patterns in the order of high evaluation as the following:

$$(0,0,0,0,1) \quad (0,0,0,1,0) \quad (0,0,1,0,0) \quad (0,1,0,0,0) \quad (1,0,0,0,0)$$

This fact appears to suggest that the philosophy of scoring underlying the normal ogive model is such that an examinee is evaluated as low in the ability tested because he can only solve an easy item, while an examinee who has solved a difficult item is evaluated as high in the ability tested because of the difficulty of the item solved.

Now let us consider another instance, if there are another five examinees who have also tried to solve the same five items and who have succeeded in four of them, but failed in only one, and each item failed is different from each other, to which of these five examinees should be assigned the highest value of estimate? On the normal ogive model, again we can easily see that these five response patterns are arranged in the order of the magnitudes of estimates as follows:

$$(1,1,1,1,0) \quad (1,1,1,0,1) \quad (1,1,0,1,1) \quad (1,0,1,1,1) \quad (0,1,1,1,1)$$

In this case the philosophy of scoring seems to be that an examinee is evaluated as low in the ability tested because of the fact that he cannot solve even an easy item, while an examinee who has failed in a difficult item is evaluated as high in the ability tested because it is no proof of his inferiority that he has failed in a difficult item.

The above two philosophies are, in one sense, contradictory with each other, since the principle is completely reversed. That is to say, the difficulty of an item in the former instance is treated just as the easiness of an item in the latter instance, and vica versa. If we apply the principle in the former case to the latter, the order of evaluation of the five response patterns should be reversed, since, for instance, an examinee with response pattern $(0,1,1,1,1)$ has succeeded in the most difficult four items, while an examinee with response pattern $(1,1,1,1,0)$ has succeeded in the easiest four items."

A solution of this contradiction can be found in the adoption of a certain asymmetric model. Samejima (1972) called the family of models represented by

$$P_g(\theta) = \left[\frac{1}{1+\exp[Da_g(\theta - b_g)]}\right]^{\xi_g} \tag{14}$$

with $\xi_g\ (>0)$ as the third parameter the positive exponent family of the logistic model. It can be shown that, when $\xi_g > 1$, the model consistently follows the philosophy of giving more credit to the success in answering a more difficult item correctly, providing the reversed order of $\hat{\theta}_v$'s for the response patterns that have only one item score 0 each while the order of the $\hat{\theta}_v$'s for the response patterns having only one item score 1 each is unchanged. This is illustrated in Table 1 with $n = 5$, and with the parameter values $a_g = 1$ for all items, $b_g = -3.0, -1.5, 0.0, 1.5, 3.0$, respectively, and $\xi_g = 2$.

Insert Table 1 About Here

In this situation, rationale behind the family of models represented by (14) includes the *slip ratio* defined as an monotone decreasing function of θ. When it is incorporated in the logistic model represented by (11), if the slip function is assumed to be

$$1 - \frac{1}{1+\exp[Da_g(\theta - b_g)]}$$

which is strictly decreasing in θ, the third parameter ξ_g in (14) will become 2; if it is assumed to be

$$1 - \left[\frac{1}{1+\exp[Da_g(\theta - b_g)]}\right]^{\frac{1}{2}}$$

which is also strictly decreasing in θ, ξ_g will become 1.5 in (14), etc.

When $\xi_g < 1$, the model represented by (14) follows the principle that failure in answering an easy item correctly be penalized more than failure in answering a more dif-

ficult item correctly. Rationale behind this family of models, including both of the above two situations, is discussed in detail in a separate paper (Samejima, in preparation).

IV. Discussion and Conclusions

It should be kept in mind that the objective of psychological measurement is to estimate the latent trait as accurately as possible making the best use of information provided by item responses. The use of the test score instead of the response pattern itself in ability or attitude estimation will, in general, reduce accuracy of estimation, and the loss of accuracy can be enormous especially when items are scored polychotomously. It is strongly suggested that ability or attitude estimation be made directly from the response pattern using basic functions given by (10), instead of using the test score as the estimate of ability or attitude, or estimating the latent trait from the test score.

In so doing, substantive model validation is essential, for if we use an inappropriate model for our data information provided by the model will be useless, or even harmful. Usefulness of models that provide asymmetric operating characteristics for dichotomous items, or its expansion to ordered polychotomous items such as the acceleration model (Samejima, 1995) should seriously be considered.

References

[1] Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, Contributed chapters in Lord, F. M. and Novick, M. R., *Statistical theories of mental test scores*, Chapters 17-20. Reading, MA: Addison Wesley.

[2] Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

[3] Grayson, D. A. (1988). Two group classification in latent trait theory: Scores with

monotone likelihood ratio. *Psychometrika, 53*, 283-292.

[4] Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59*, 77-79.

[5] Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

[6] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielson & Lydiche.

[7] Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

[8] Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No. 18.*

[9] Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory, *Psychometrika, 38*, 221-233.

[10] Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika, 60*, 549-572.

[11] Samejima, F. (1996). Acceleration model in the heterogeneous case of the general graded response model. *Behaviormetrika*, (in press).

[12] Samejima, F. (1996). Virtue of asymmetric item charactristic curves. (in preparation).

[13] Yen, W. M., Burket, G. R. and Sykes, R. C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model, *Psychometrika, 56*, 39-54.

NCME962.TEX

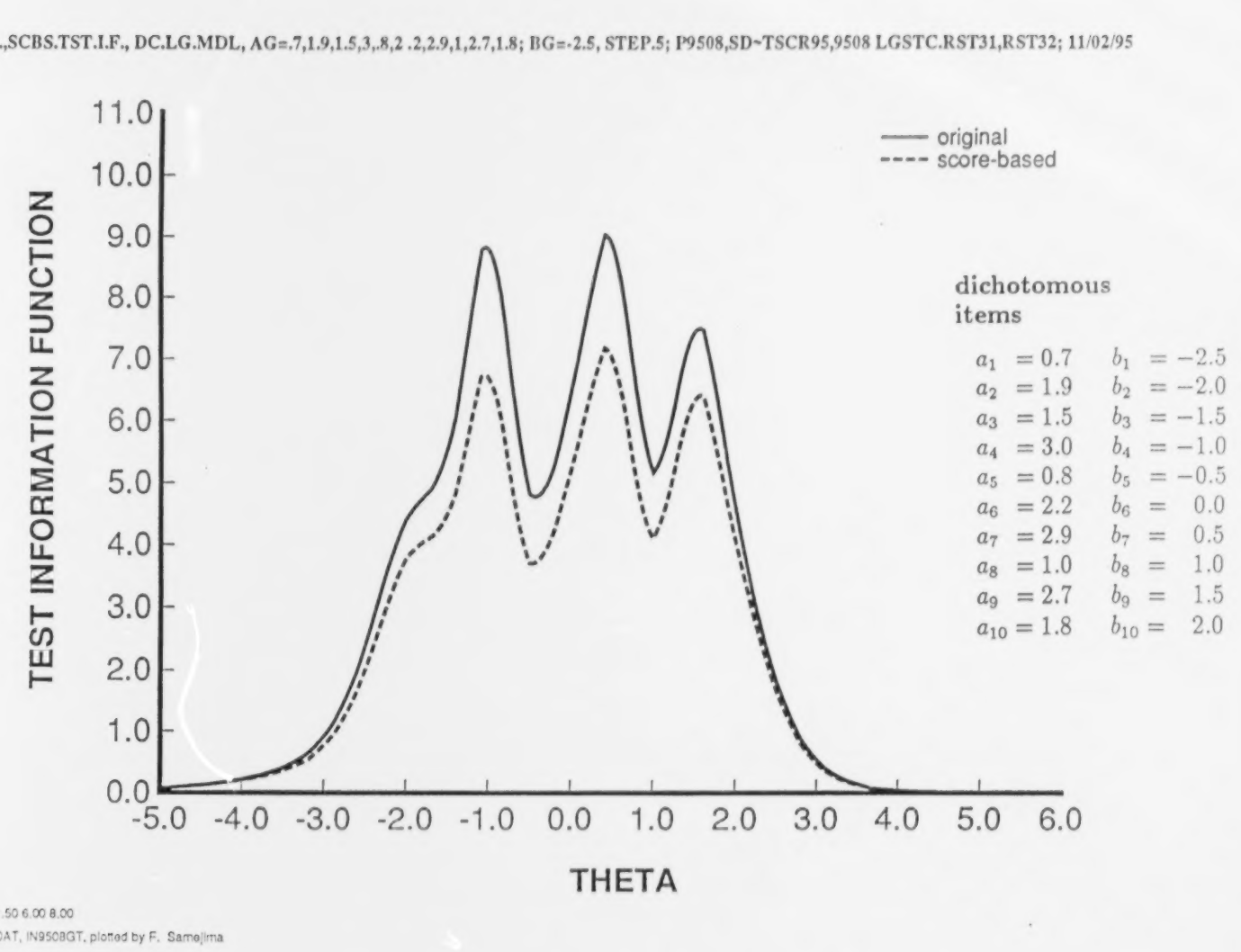

FIGURE 1

The original test information function and the one based on the test score of a hypothetcal test of 10 dichotomous items following the logistic model.

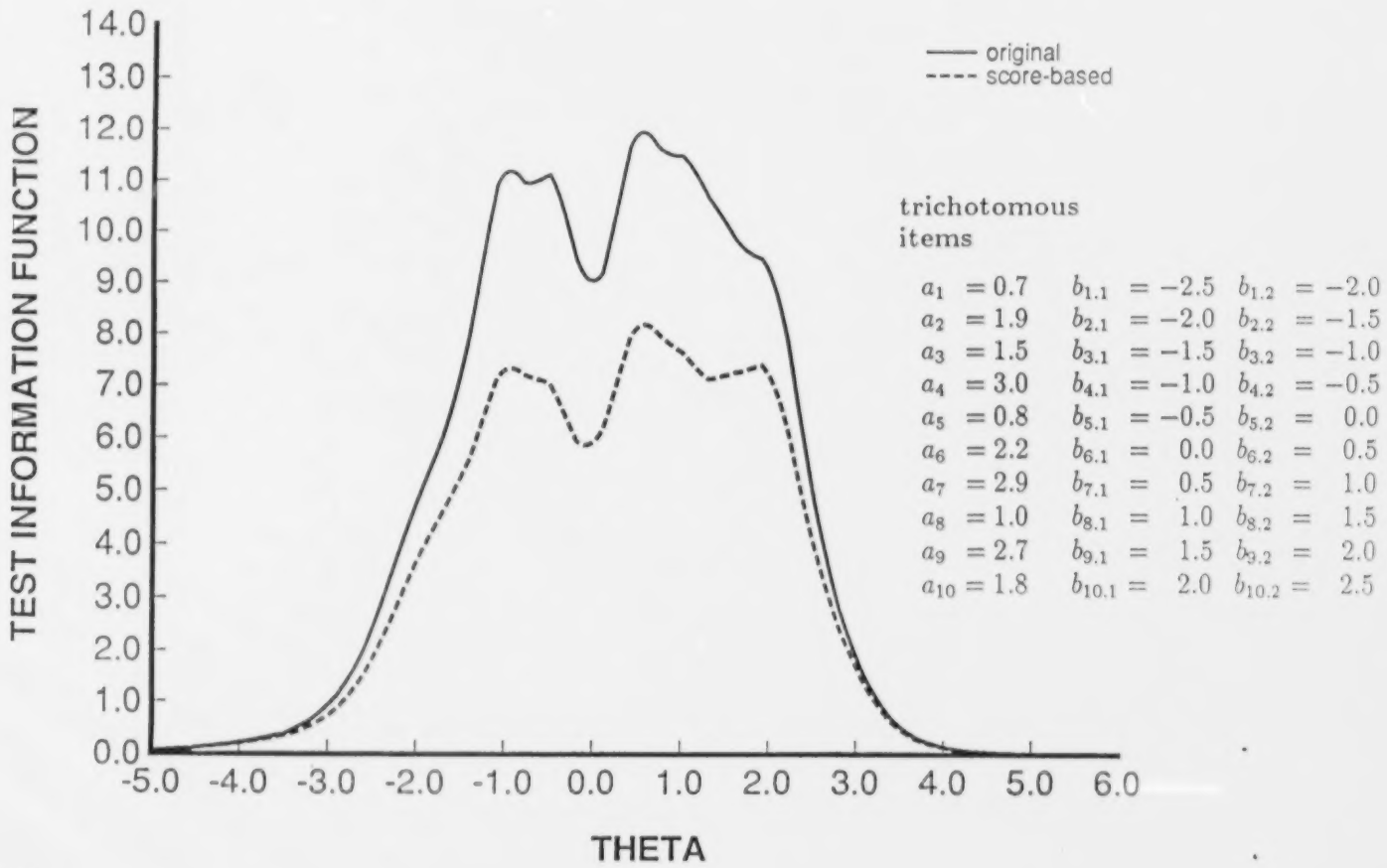

FIGURE 2

The original test information function and the one based on the test score of a hypothetcal test of 10 graded items with $m_g = 2$, following the logistic model.

TABLE 1

The Maximum Likelihood Estimates of θ Based on 32 Response Patterns of 5 Dichotomous Items Following the Normal Ogive Model with the Parameters $a_g = 1.0$ for All Items and $b_g = -3.0, -1.5, 0.0, 1.5, 3.0$, Arranged in the Ascending Order.

	Response Pattern	MLE Theta
1	00000	neg.infinity
2	10000	-1.77109
3	01000	-1.40646
4	00100	-0.83936
5	00010	-0.44235
6	00001	-0.34778
7	11000	-0.24612
8	10100	0.10334
9	01100	0.13035
10	10010	0.66449
11	01010	0.67548
12	00110	0.77745
13	10001	1.06032
14	01001	1.06796
15	00101	1.14590
16	11100	1.25580
17	00011	1.47795
18	11010	1.60421
19	10110	1.63116
20	01110	1.63323
21	11001	2.16546
22	10101	2.17644
23	01101	2.17729
24	10011	2.27846
25	01011	2.27907
26	00111	2.28672
27	11110	2.76207
28	11101	3.11779
29	11011	3.14533
30	10111	3.14744
31	01111	3.14760
32	11111	pos.infinity